AD-A114 070  MASSACHUSETTS INST OF TECH  LEXINGTON LINCOLN LAB  F/G 17/2
MAXIMUM LIKELIHOOD SPECTRAL ESTIMATION AND ITS APPLICATION TO N--ETC(U)
MAR 82  R J MCAULAY  F19628-80-C-0002
UNCLASSIFIED  TR-602  ESD-TR-82-006  NL

1 OF 1
ADA
114070

END
DATE
FILMED
5-82
DTIC

# Technical Report

602

R.J. McAulay

## Maximum Likelihood Spectral Estimation and Its Application to Narrowband Speech Coding

5 March [illegible]

[illegible]

Lincoln Laboratory

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

LEXINGTON, MASSACHUSETTS

[illegible]

MASSACHUSETTS INSTITUTE OF TECHNOLOGY
LINCOLN LABORATORY

# MAXIMUM LIKELIHOOD SPECTRAL ESTIMATION AND ITS APPLICATION TO NARROWBAND SPEECH CODING

*R.J. McAULAY*
*Group 24*

TECHNICAL REPORT 602

5 MARCH 1982

Approved for public release; distribution unlimited.

LEXINGTON MASSACHUSETTS

# ABSTRACT

The maximum likelihood (ML) method has been used by Itakura and Saito [1] to derive a nonlinear spectral matching criterion for estimating the spectral parameters of autoregressive (AR) processes. In this paper it is shown that their spectral matching criterion is a general property of ML spectral estimation in that it is valid for any spectral model and applies to aperiodic <u>and</u> periodic random processes.

An exact solution to the ML parameter estimation problem for AR processes has recently been derived by Kay [2]. These results are cast in a frequency domain formulation which is used to generalize the exact solution to periodic processes. It is then shown that if the number of independent power measurements, N, greatly exceeds the model order, M, then the ML algorithm reduces to a pitch-directed frequency domain version of Linear Predictive (LP) spectral analysis.

The exact solution is then used to determine the spectral envelope for voiced (periodic) and unvoiced (aperiodic) speech and it is observed that the exact analysis results in fits that broaden the formant bandwidths while reducing the formant amplitudes. A real-time vocoder was developed and it was found that in contrast to a standard LPC algorithm the exact ML analysis produced synthetic speech that had the quality of being heavily smoothed. This perceptual difference shows that it is generally incorrect to interpret LPC spectral matching in terms of the Itakura-Saito criterion.

# CONTENTS

## I. INTRODUCTION AND SUMMARY

Itakura and Saito [1] have shown that spectral envelope estimation using linear predictive coding techniques (LPC) has a more fundamental theoretical basis in maximum likelihood (ML) estimation. Furthermore, they have used this theory to develop a spectral matching interpretation in terms of the Itakura-Saito criterion. Their basic mathematical model dealt with speech waveforms that were sample functions of an autoregressive random process for which the spectrum was not harmonic. While this is the correct model for the class of unvoiced sounds, one wonders if perhaps the results are valid for voiced speech sounds as well, since in this case the waveforms are periodic with spectra having distinct harmonic line structure. This is the problem addressed in this paper.

In setting up the formalism for the application of the ML method to both aperiodic and periodic processes, it was not necessary to impose the all-pole constraint on the model spectrum. The ensuing analysis led to a spectral matching criterion identical to that obtained by Itakura and Saito, which shows that the criterion is a fundamental property of the maximum likelihood method. Since this is a deterministic function of the measured and model spectra, the underlying Gaussian statistical model is no longer a significant limitation to the ultimate validity of the results and, for all intents and purposes, can be ignored. Furthermore, this interpretation shows that in the periodic case, the model spectrum is fitted to the power measurements at the harmonic frequencies.

In a recent paper Kay [2] obtained an exact expression for the ML estimates of the parameters of an autoregressive process using a covariance domain analysis. By modifying Kay's analysis to reflect spectral domain properties, the ML formalism derived in this paper could be used to obtain an exact solution for the parameters of an all-pole spectral envelope for aperiodic and periodic processes. It was shown that if the number of independent power measurements, N, (the number of pitch harmonics in the periodic case) greatly exceeds the model order, M, then the ML algorithm reduced to a pitch-directed frequency domain version of linear prediction (LP) spectral analysis. In this case the fundamental measurement set is not a set of correlation coefficients, but a set of power spectrum measurements.

The exact ML solution was then used to determine the spectral envelope for voiced (periodic) and unvoiced (aperiodic) speech. It was observed that for voiced speech the exact method led to a "less faithful" reproduction of the spectral measurements and resulted in broadened formant bandwidths and reduced formant amplitudes. During unvoiced speech the spectral fits were in general agreement, which is consistent with the theoretical results. In order to determine whether or not these differences were perceptually significant a real-time analysis/synthesis system was developed. It was found that the synthetic speech produced by the exact ML algorithm had the quality of being too heavily smoothed. Although this had the effect of eliminating small spectral distortions which occasionally occurred in the approximate (N>>M) analysis, the approximate system produced synthetic speech which was more natural. Furthermore, when compared with a standard autocorrelation based LPC vocoder using the same acoustic tube synthesizer, the pitch-directed frequency based approximate ML algorithm did not result in synthetic speech that was significantly better either in quality or intelligibility.

Based on the experience to date, it appears that the major benefit of the exact ML analysis for aperiodic and periodic processes having all-pole spectral envelopes, is the fact that it leads to a unified theoretical formulation for analyzing voiced and unvoiced speech. It turns out, however, that the frequency domain implementation of the spectral analysis algorithm has been found to be particularly useful in the development of low rate systems [3], and higher quality split-band vocoder algorithms.

## II. THEORETICAL FOUNDATION

In order to provide a common theoretical framework for estimating the spectra for voiced (periodic) and unvoiced (aperiodic) speech, it is useful to model the speech waveform S(n) as a sample function of a suitably defined discrete time random process. A key step in establishing the analytical framework is to expand S(n) in terms of a set of basis functions in such a way that the expansion coefficients are uncorrelated random variables. For unvoiced speech, the theory requires that the sample functions be defined on a finite interval N points long. A particularly important set of basis functions are the complex exponentials, for which the series expansion is

$$S(n) = \sum_{k=o}^{N-1} X_k \exp\,(jn\omega_k) \qquad (1)$$

where $\omega_k = 2\pi k/N$ and

$$X_k = \frac{1}{N} \sum_{n=o}^{N-1} S(n) \exp\,(-jn\omega_k) \qquad (2)$$

It is useful to determine the conditions under which the complex exponentials can be considered to be eigenfunctions of the covariance matrix $R(n,m) = E[S(n)S^*(m)]$. This occurs when

$$\lambda_k \exp\,(jn\omega_k) = \sum_{m=o}^{N-1} R(n,m) \exp\,(jm\omega_k) \qquad (3)$$

for some value of $\lambda_k$. If the unvoiced speech process is wide sense stationary then

$$R(n,m) = R(n-m)$$

$$= \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega) \exp [j\omega(n-m)]d\omega \qquad (4)$$

where $P(\omega)$ represents the power spectral density. Substituting (4) into (3) leads to the eigenvalue equation

$$\lambda_k \cdot \exp (jn\omega_k) = \frac{1}{2\pi}\int_{-\pi}^{\pi} P(\omega) \exp (jn\omega)\ g(\omega_k-\omega)d\omega \qquad (5)$$

where

$$g(\theta) = \sum_{m=o}^{N-1} \exp (jm\theta)$$

$$= \exp[j(N-1)\theta/2]\cdot\frac{\sin(N\theta/2)}{\sin(\theta/2)} \qquad (6)$$

If the frame length N is large enough that the power spectral density changes slowlv in a frequency increment 1/N, then relative to the function $P(\omega)$, $g(\omega_k-\omega)$ can be considered to be an impulse at $\omega_k$. Under these conditions (5) reduces to

$$\lambda_k \exp (jn\omega_k) \approx \frac{1}{2\pi} P(\omega_k) \exp (jn\omega_k) \qquad (7)$$

which shows that the complex exponentials are valid eigenfunctions of $R(n,m)$. The associated eigenvalues are

$$\lambda_k = \frac{1}{2\pi}P(\omega_k) \qquad (8)$$

Finally, using (2) and (3) it is easy to show that the correlation between expansion coefficients is given by

$$E(X_k X_i^*) = \frac{1}{2\pi N} P(\omega_k)\delta_{ki} \qquad (9)$$

the desired result.

For voiced speech the analysis presumes that the random process S(n) is periodic where N now represents the period. This means that the covariance function $R(m) = E[S(n)S^*(n+m)]$ is periodic with period N. Although this model has no relevance to the physiological mechanism by which voiced sounds are generated, mathematically it can be used to generate a class of spectra that have roughly the same properties as voiced speech spectra, and hence it was adopted as being suitable for the type of analysis to be undertaken. Now if the covariance function is periodic, then $R(m) = R(m+N)$ for all m, and as a consequence can be expanded as

$$R(m) = \sum_{k=o}^{N-1} P_k \exp\ (jk\omega_m) \qquad (10)$$

where now $\omega_m = 2\pi m/N$, and where

$$P_k = \sum_{m=o}^{N-1} R(m) \exp\ (-jk\omega_m) \qquad (11)$$

which specifies the discrete time power spectrum of the voiced speech process. A series representation for the original random process can be taken to be

$$S(n) = \sum_{k=o}^{N-1} X_k \exp\ (jn\omega_k) \qquad (12)$$

where

$$X_k = \frac{1}{N} \sum_{n=o}^{N-1} S(n) \exp\ (-jn\omega_k) \qquad (13)$$

It is easy to show that the correlation between the voiced speech expansion coefficients is given by

$$E(X_k X_i^*) = P_k \delta_{ki} \tag{14}$$

Therefore, both voiced (periodic) and unvoiced (aperiodic) speech can be expanded in terms of a set of uncorrelated random variables with respect to the complex exponential basis. For unvoiced speech the basis was defined on an interval N such that in a frequency increment 1/N, the unvoiced power spectral density was slowly changing. For voiced speech the basis was defined on an interval N that was simply one period of the periodic covariance function. The spectrum estimation problem for the voiced and unvoiced speech cases can be specified in terms of a common parameter estimation problem by defining the "spectrum" by

$$\lambda_k(\underline{\theta}) = \begin{cases} P_k & \text{voiced speech} \\ \frac{1}{2\pi N} P(\omega_k) & \text{unvoiced speech} \end{cases} \tag{15}$$

Although the functional form of $\lambda_k(\underline{\theta})$ may be known, the parameters, $\underline{\theta}$, upon which that form depends are not. The problem is to use the set of measurements $\{X_k\}$ in (2) or (13) to produce the best possible estimate of the spectral parameters. The maximum likelihood (ML) method can lead to estimates that are known to have several asymptotic optimality properties [4]. Computation of the ML estimates requires specification of a probabilistic model for the speech processes and mathematical tractability usually requires that a Gaussian model be used, which can be justified by the usefulness of the results it produces. Henceforth, the random variables $\{X_n\}$ will be assumed to have a Gaussian distribution with zero mean. Since these random variables have already been shown to be uncorrelated, the Gaussian assumption implies that they are also independent. The joint probability density function (pdf) can therefore be written explicitly as

$$P(X_1,X_2,\ldots,X_N|\underline{\theta}) = \prod_{n=o}^{N-1}\left\{\frac{1}{\pi\lambda_n(\underline{\theta})}\exp\left[-\frac{|X_n|^2}{\lambda_n(\underline{\theta})}\right]\right\}$$

$$= \pi^{-N}\exp\left\{-\sum_{n=o}^{N-1}\left[\frac{|X_n|^2}{\lambda_n(\underline{\theta})} + \log\lambda_n(\underline{\theta})\right]\right\} \quad (16)$$

The ML estimates of $\underline{\theta}$ are found by maximizing this pdf. This is equivalent to minimizing the negative of the logarithm of the pdf which is called the likelihood function and is written as

$$\ell(\underline{\theta}) \overset{\Delta}{=} -\log\left[P(X_1,X_2,\ldots X_N|\underline{\theta})\right]$$

$$= N\log\pi + \sum_{n=o}^{N-1}\left[\frac{|X_n|^2}{\lambda_n(\underline{\theta})} + \log\lambda_n(\underline{\theta})\right] \quad (17)$$

From this equation the ideas first proposed by Itakura and Saito [1] can be used to develop a spectral matching interpretation of the ML criterion. The first step is to obtain a lower bound on the likelihood function by obtaining the ML estimates for the unconstrained problem. For this case there are N unknown parameters $\lambda_n$ and the ML estimates are easily shown to be $\hat{\lambda}_n = |X_n|^2$. The resulting minimum value of the likelihood function is

$$\ell_{min} = N\log\pi + \sum_{n=o}^{N-1}(1+\log|X_n|^2) \quad (18)$$

The likelihood function at any other value can then be written as

$$\ell(\underline{\theta}) - \ell_{min} = \sum_{n=o}^{N-1}\left\{\exp\left[\log\left(\frac{|X_n|^2}{\lambda_n(\underline{\theta})}\right)\right] - \log\left(\frac{|X_n|^2}{\lambda_n(\underline{\theta})}\right) - 1\right\} \quad (19)$$

The next step is to define the quantity

$$E(f_n) = \log|X_n|^2 - \log\lambda_n(\underline{\theta}) \quad (20)$$

which measures the dB error between the power at frequency $f_n$ measured by $|X_n|^2$ and estimated by the model as $\lambda_n(\underline{\theta})$. Using this, the ML criterion in (19) can be expressed by the nonlinear spectral matching condition

$$\ell(\underline{\theta}) - \ell_{min} = \sum_{n=o}^{N-1} \left\{ \exp\left[E(f_n)\right] - E(f_n) - 1 \right\} \tag{21}$$

Following [1] it is of interest to contrast this result with that obtained with the squared dB error criterion given by

$$f(\underline{\theta}) = \sum_{n=o}^{N-1} [E(f_n)]^2 \tag{22}$$

As shown in Fig. 1, this condition gives equal weight to dB model errors above and below the measured power samples, whereas the ML criterion gives significantly more weight to errors that occur when the model spectrum lies below the measured data. This suggests that the ML parameter estimates will result in a model spectrum that "sits on top of" the spectral measurements, hence it would appear to be well suited for estimating a spectral envelope from discrete spectral measurements.

Although these properties are well-known in the speech community, they were thought to apply only to the case of aperiodic speech which could be modelled in terms of an autoregressive (all-pole) process. The significance of the above analysis is that the nonlinear spectral matching interpretation applies to voiced and unvoiced speech and does not depend on a specific spectral model for the speech generation process. Hereafter, we shall refer to (21) as the maximum likelihood spectral matching criterion.

Another advantage of the maximum likelihood formulation is the insight it provides into the way in which the spectral measurements should be made. For unvoiced speech the measurement variables are

$$X_k = \frac{1}{N} \sum_{n=o}^{N-1} S(n) \exp\left(\frac{-j2\pi nk}{N}\right) \tag{23}$$

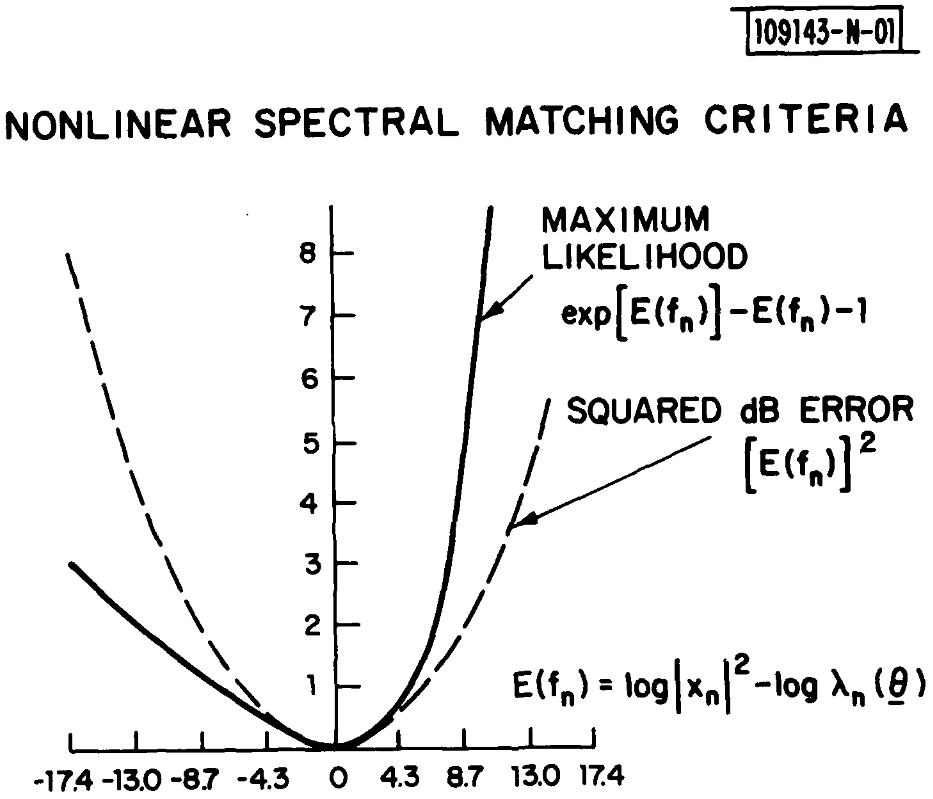

Fig. 1. Maximum likelihood spectral matching criterion.

which is simply the Discrete Fourier Transform (DFT) of the speech process sampled at frequency n/N. The data satisfy the conditions of the theory provided the frame length N is large enough that the actual power spectral density changes slowly in a frequency increment 1/N. Since it is standard practice in vocoder design to choose an analysis window of at least 20 ms and since typical unvoiced speech spectra change very little in 50 Hz, then the conditions of the theory are met in practice. For voiced speech the measurements are

$$X_k = \frac{1}{N} \sum_{n=o}^{N-1} S(n) \exp\left(\frac{-j2\pi nk}{N}\right) \tag{24}$$

where N is the period of the voiced speech process. This result suggests that the pitch must be known and used explicitly in developing the correct measurement set for computing the ML estimates, a procedure which is not commonly used in practice. This is an extremely important point because it suggests that if the data truly correspond to a periodic process, then simply assuming that the speech can be modelled as an AR process is not sufficient for optimally extracting the measurement variables. This issue will be reexamined later in the sequel after the solution to the all-pole modelling problem has been derived.

## III. ALL-POLE MODELLING

Although the ML criterion that was derived in the previous section can be used to estimate the parameters of model spectra of arbitrary functional form, there is particular interest in the all-pole model since it corresponds to autoregressive (AR) processes that arise in diverse applications of spectral analysis, and, in speech in particular, it is related to the Linear Prediction method of speech coding. Recently Kay [2] has developed a recursive solution to the exact ML problem for aperiodic AR processes using a covariance domain solution technique that is an extension of the covariance method of Linear Prediction. In this section Kay's analysis will be generalized in terms of the spectral domain formulation of the ML problem that was developed in the previous section, and will thereby yield the exact ML spectrum for aperiodic and periodic processes.

The analysis begins with a restatement of the ML estimation problem for processes for which the spectral eigenvalues have the all-pole form:

$$\lambda_n(\underline{\theta}) = \frac{\sigma_M^2}{|A_M(\omega_n)|^2} \tag{25}$$

where, as before, $\omega_n = 2\pi n/N$ and where M is the model order and

$$A_M(\omega_n) = 1 - \sum_{m=1}^{M} a_m^{(M)} \exp(-jm\omega_n) \tag{26}$$

The parameters to be estimated are

$$\underline{\theta} = \left[\sigma_M^2, a_1^{(M)}, a_2^{(M)}, \ldots, a_M^{(M)}\right] \tag{27}$$

The ML estimate of $\underline{\theta}$ is obtained by minimizing the ML spectral matching criterion in (19) or equivalently the likelihood function in (17) which is

$$\ell(\underline{\theta}) = \sum_{n=o}^{N-1} \frac{|X_n|^2}{\lambda_n(\underline{\theta})} + \sum_{n=o}^{N-1} \log \lambda_n(\underline{\theta}) + N \log \pi \tag{28}$$

McAulay [10] has shown that the key to deriving the exact ML solution depends on the derivation of an alternate expression for the second term in (28). As a first step it is obvious that

$$\begin{aligned}\sum_{n=o}^{N-1} \log \lambda_n &= \log\left(\prod_{n=o}^{N-1} \lambda_n\right) \\ &= \log\left(\det \Lambda\right)\end{aligned} \tag{29}$$

where $\Lambda$ is the NxN diagonal matrix

$$\Lambda = \text{diag}\ (\lambda_o, \lambda_1, \ldots, \lambda_{N-1}) \tag{30}$$

Hence, the problem reduces to finding an equivalent expression for this matrix. This can be done by noting that the underlying model covariance matrix satisfies the relations

$$\begin{aligned} R(t,s) = E\left[S(t)S^*(s)\right] &= E\left[\sum_{i=o}^{N-1} X_i\phi_i(t) \sum_{j=o}^{N-1} X_j^*\phi_j^*(s)\right] \\ &= \sum_{n=o}^{N-1} \lambda_n\phi_n(t)\ \phi_n^*(s) \end{aligned} \tag{31}$$

This is known as Mercer's Theorem and follows from the fact that $E(X_iX_j^*) = \lambda_i\delta_{ij}$, a condition which was derived in Section II. Equation (31) can be written in matrix notation as

$$\underline{R} = \Phi^T\Lambda\Phi^* \tag{32}$$

by defining the NxN covariance matrix $(\underline{R})_{ij} = R(i,j)$ and the NxN transformation matrix $\Phi$ as

$$\Phi = \begin{bmatrix} \phi_o(o) & \phi_o(1) & \ldots & \phi_o(N-1) \\ \phi_1(o) & \phi_1(1) & \ldots & \phi_1(N-1) \\ \vdots & \vdots & & \\ \phi_{N-1}(o) & \phi_{N-1}(1) & \ldots & \phi_{N-1}(N-1) \end{bmatrix} \tag{33}$$

where $\Phi^T$ and $\Phi^*$ are the transpose and the conjugate of the matrix $\Phi$ respectively. As a consequence of (32)

$$\det(R) = \det(\Phi^T\Phi^*).\ \det\ (\Lambda) \tag{34}$$

However, the elements of $\Phi$ are the expansion eigenfunctions which in this case are the complex exponentials, viz $\phi_n\ (k) = \exp\ (-j2\pi nk/N)$, and is therefore related to the Discrete Fourier Transform (DFT). In fact the DFT of a vector $\underline{z}$ is

$$\underline{Z} \triangleq \text{DFT}\ (\underline{z}) = \Phi\ \underline{z} \tag{35}$$

hence, the Inverse DFT is

$$\underline{z} = \frac{1}{N}\ \Phi^*\ \underline{Z} \tag{36}$$

which implies that $\frac{1}{N}\ \Phi^*\ \Phi = I$, the identity matrix, from which it follows that $\det\ (\Phi^T \Phi^*) = N$. Using this result in (34) leads to the equation

$$\det\ (\Lambda) = \frac{1}{N} \det\ (\underline{R}) \tag{37}$$

hence the exact solution to the ML problem depends on the evaluation of det $(\underline{R})$. This latter computation has been done by Kay [2] in a covariance domain derivation of the exact solution to the ML problem. He showed that

$$\det\ (\underline{R}) = \frac{\sigma_M^{2N}}{\prod_{m=1}^{M} (1-K_m^2)^m} \tag{38}$$

where $K_1, K_2, \ldots, K_M$ are the so-called reflection coefficients and are related to the original all-pole spectral model through the Levinson-Durbin algorithm [6] which requires that

$$a_i^{(m)} = \begin{cases} a_i^{(m-1)} - K_m\ a_{m-i}^{(m-1)} & i=1,2,\ldots,m-1 \\ K_m & i=m \end{cases} \tag{39}$$

As a consequence of (29), (37) and (38) it follows that

$$\sum_{n=o}^{N-1} \log \lambda_n = N \log \sigma_M^2 - \sum_{m=1}^{M} m \log\ (1-K_m^2) - \log N \tag{40}$$

which can be substituted into (28) to result in the following equation for the exact likelihood function:

$$\ell(\underline{\theta}) = \frac{1}{\sigma_M^2} \sum_{n=o}^{N-1} |X_n|^2 |A_M(\omega_n)|^2 + N \log \sigma_M^2 - \sum_{m=1}^{M} m \log (1-K_m^2) + \log\left(\left[\frac{(\pi^N)}{N}\right]\right) \quad (41)$$

Since equation (39) can be used with (26) to derive the well-known recursions:

$$A_m(\omega) = A_{m-1}(\omega) + K_m B_{m-1}(\omega) \quad (42a)$$

$$B_m(\omega) = \exp(-j\omega)\ [B_{m-1}(\omega) + K_m A_{m-1}(\omega)] \quad (42b)$$

which are initialized at stage m=o with the conditions

$$A_o(\omega) = 1 \quad (43a)$$

$$B_o(\omega) = -\exp(-j\omega) \quad (43b)$$

it follows that the likelihood function now depends only on the parameters $\sigma_M^2$, $K_1$, $K_2$, ..., $K_M$. Equations (41)-(43) represent the spectral domain equivalent of Kay's covariance based expression for the exact likelihood function. However, by suitably interpreting the meaning of the power measurements, $|X_n|^2$, the above results apply to the more general case that includes aperiodic <u>and</u> periodic processes.

The gain optimized likelihood function is obtained by choosing $\sigma_M^2$ such that $\partial\ell/\partial\sigma_M^2 = 0$. This results in the ML estimate for the gain which is

$$\hat{\sigma}_M^2 = \frac{1}{N} \sum_{n=o}^{N-1} |X_n|^2 |A_M(\omega)|^2 \quad (44)$$

and the gain optimized likelihood function becomes

$$\ell(K_1,K_2,\ldots,K_M) = N \log \hat{\sigma}_M^2 - \sum_{m=1}^{M} m \log (1-K_m^2) + \log \frac{(\pi e)^N}{N} \quad (45)$$

In order to derive a simple recursion for $\ell(\underline{K})$ the vectors

$$\underline{A}_M = \Big(A_M(\omega_o),\ A_M(\omega_1),\ \ldots,\ A_M(\omega_{N-1})\Big) \tag{46a}$$

$$\underline{B}_M = \Big(B_M(\omega_o),\ B_M(\omega_1),\ \ldots,\ B_M(\omega_{N-1})\Big) \tag{46b}$$

are used to define an inner product

$$<\underline{A}_M,\ \underline{B}_M> = \frac{1}{N}\sum_{n=0}^{N-1} |X_n|^2\ A_M(\omega_n)B_M^*(\omega_n) \tag{47}$$

By defining the variables

$$\alpha_M = \mathrm{Re} <\underline{A}_M,\ \underline{B}_M> \tag{48a}$$

$$\beta_M = \|\underline{B}_M\|^2 \tag{48b}$$

and using the recursion in (42a) for $A_M(\omega)$ it follows that

$$\hat{\sigma}_M^2 = \|\underline{A}_M\|^2$$

$$= \hat{\sigma}_{M-1}^2 + 2K_M\ \alpha_{M-1} + K_M^2\ \beta_{M-1} \tag{49}$$

However, the recursion in (42b) results in the fact that

$$\beta_M = \beta_{M-1} + 2K_M\ \alpha_{M-1} + K_M^2\ \hat{\sigma}_{M-1}^2 \tag{50}$$

and the initial conditions in (43) require that

$$\hat{\sigma}_o^2 = \frac{1}{N}\sum_{n=o}^{N-1} |X_n|^2 = \beta_o \tag{51}$$

As a consequence $\hat{\sigma}_M^2$ and $\beta_M$ obey the same recursion, and therefore

$$\hat{\sigma}_M^2 = \beta_M \tag{52}$$

for all M. As a result the recursion for $\hat{\sigma}_M^2$ can be written as

$$\hat{\sigma}_M^2 = \hat{\sigma}_{M-1}^2\ (1 + 2\ \rho_{M-1}\ K_M + K_M^2) \tag{53}$$

where

$$\rho_{M-1} = \frac{\alpha_{M-1}}{\hat{\sigma}^2_{M-1}} \tag{54}$$

Letting $\ell_M = \ell(K_1, K_2, \ldots, K_M)$, and substituting (53) in (45) leads to the expression

$$\ell_M = N \log \hat{\sigma}^2_{M-1} + N \log (1 + 2\rho_{M-1}K_M+K_M^2) - \sum_{m=1}^{M} m \log (1-K_m^2) + \log \left[\frac{(\pi e)^N}{N}\right]$$

$$= \ell_{M-1} + N \log (1+2\rho_{M-1}K_M+K_M^2) - M \log (1-K_M^2) \tag{55}$$

which with the initial condition

$$\ell_o = N \log \hat{\sigma}^2_o + \log \left[\frac{(\pi e)^N}{N}\right] \tag{56}$$

is the desired recursion. This is a somewhat simpler expression than that obtained by Kay due to the fact that the frequency domain formulation led to the condition $\hat{\sigma}^2_M = \beta_M$.

When Itakura and Saito studied the problem of ML estimation of the spectral parameters of all-pole aperiodic processes, they invoked the condition that N, the number of independent spectral measurements, greatly exceeded the model order, M, (i.e., N>>M). Under this condition the likelihood function in (55) reduces to

$$\ell_M = \ell_{M-1} + N \log ( 1 + 2\rho_{M-1} K_M + K_M^2) \tag{57}$$

If the optimum values for $K_1, K_2, \ldots, K_{M-1}$ have been found, then $\ell_{M-1}$ takes its smallest value and $\ell_M$ is minimized simply by choosing $\hat{K}_M$ to satisfy $\partial \ell_M / \partial K_M = 0$. This results in the estimator

$$\hat{K}_M = - \rho_{M-1} = - \frac{\alpha_{M-1}}{\hat{\sigma}_{M-1}} \tag{58}$$

The recursion for the spectral gain, equation (53), then simplifies to

$$\hat{\sigma}^2_M = \hat{\sigma}^2_{M-1} (1-\hat{K}^2_M) \tag{59}$$

Furthermore, application of the Schwartz Inequality to (48a) is sufficient to show that the ML reflection coefficients will satisfy the condition $|K_M| < 1$. Therefore, the results for the condition $N >> M$ are consistent with those that have been obtained by Markel and Gray [5] using the orthogonal polynomial approach to LPC analysis.

For the more general case for which the values of N and M are arbitrary, the ML estimate for $K_M$ must also satisfy the condition $\partial \ell_M / \partial K_M = 0$, which leads to the cubic equation

$$(N-M)\ \hat{K}_M^3 - (N-2M)\ \rho_{M-1}\ \hat{K}_M^2 - (N+M)\ \hat{K}_M - N\rho_{M-1} = 0 \qquad (60)$$

Although there are three roots to this equation, inspection of (55) shows that there will be only one root that satisfies $|K_M| < 1$. Since closed form expressions are available for the roots of equation (60), then, if computations are being done using floating point arithmetic, it is straightforward to determine the ML estimator numerically. In speech applications which require real-time processing using fixed point arithmetic, it has been found that Newton-Raphson methods are easier to implement. In the next section some examples will be given that show the difference in the spectral estimates obtained using the exact and approximate ML estimators.

## IV. APPLICATION TO SPEECH CODING

In contrast to the standard approaches to all-pole spectral analysis that are based on estimates of the ensemble covariance function computed from time-domain data, the ML procedure estimates the power spectral density through the frequency domain power measurement $|X_n|^2$ where

$$X_n = \frac{1}{N} \sum_{n=o}^{N-1} S(k) \exp\ (-j2\pi nk/N) \qquad (61)$$

where N represents the frame length for aperiodic processes and the period for periodic processes. In order to make the requisite power measurements in the latter case, the pitch period must be known with precision, which is difficult to achieve in practice. In addition, evaluation of the DFT in (61) represents a computationally intensive task. A more efficient approach would be to use the Fast Fourier Transform (FFT) to generate a high resolution spectrum which could be sampled at the pitch harmonics to determine the desired power measurements. The problem is, however, that not only is it difficult to know the true pitch with precision, but in many cases the voiced speech spectrum is not always harmonic, especially above 1000 Hz. Hence, sampling the high resolution spectrum could not be expected to produce reliable estimates of the power in the pitch harmonics. In the development of the SEEVOC voice coding algorithm Paul [7] has developed a method for avoiding these problems using a heuristic that selects the largest spectral peak within successive frequency windows equal in width to the average pitch frequency. For example, if the average pitch is $\bar{\omega}_o$, then the first step is to find the largest peak in the frequency range $(\bar{\omega}_o/2, 3\bar{\omega}_o/2)$. Suppose this peak, which yields the power $|X_1|^2$, occurs at frequency $\omega_1$, then the next peak, $|X_2|^2$, is found in the region $(\omega_1+\bar{\omega}_o/2, \omega_1+3\bar{\omega}_o/2)$, and so on. If no peak exists within a particular region, then the frequency at which the largest end point occurs is chosen as the next harmonic. The algorithm has been studied extensively in simulations and in real-time implementations and has been found to produce satisfactory measurements for the harmonic powers $|X_n|^2$.

In order to reduce the computational complexity the same algorithm is used to determine the power measurements during unvoiced speech. Since the high resolution spectrum is more or less continuous (i.e., not harmonic), the peak picking algorithm essentially samples the measured spectrum at the average pitch frequency $\bar{\omega}_o$. Therefore, in summary, the power measurements $|X_n|^2$ and the frequencies $\omega_n$ that are actually used in the ML estimation procedure are taken to be those obtained from the pitch-directed peak-picking routine. Although these measurements are not harmonically related in general, the properties of the nonlinear ML spectral matching criterion described in Section II, still apply and determine the nature of the all-pole model fit to the power measurements.

The next step in determining the spectral parameters is to set up initial conditions from (43), (44), and (48a). These are

$$A_o(\omega_n) = 1 \tag{62a}$$

$$B_o\ (\omega_n) = -\exp(-j\omega_n) \tag{62b}$$

$$\hat{\sigma}_o^2 = \frac{1}{N} \sum_{n=o}^{N-1} |X_n|^2 \tag{62c}$$

$$\alpha_o = \frac{1}{N} \sum_{n=o}^{N-1} |X_n|^2 \cos \omega_n \tag{62d}$$

The first reflection coefficient is found by solving (60), which, for this case is,

$$(N-1)\ \hat{K}_1^3 + (N-2)\ \rho_o\ \hat{K}_1^2 - (N+1)\ \hat{K}_1 - N\rho_o = 0 \tag{63}$$

where $\rho_o = \alpha_o/\hat{\sigma}_o^2$. The approximate ML estimate of $K_1$ is $\hat{K}_1 = -\rho_o$, which is used to initialize a Newton-Raphson search for the solution of (63). Usually convergence to 15 bit accuracy is obtained in less than five iterations. Having obtained $\hat{K}_1$, the next step is to update the above quantities according to (42), (48a) and (53). The appropriate computations are

$$A_M(\omega_n) = A_{M-1}(\omega_n) + \hat{K}_M B_{M-1}(\omega_n) \tag{64a}$$

$$B_M(\omega_n) = \exp\ (j\omega_n)\ [B_{M-1}(\omega_n) + \hat{K}_M A_{M-1}(\omega_n)] \tag{64b}$$

$$\alpha_M = \frac{1}{N}\sum_{n=0}^{N-1}\left\{|X_n|^2 \ \mathrm{Re}\ [A_M(\omega_n)\ B^*_M(\omega_n)]\right\} \qquad (64c)$$

$$\hat{\sigma}^2_M = \hat{\sigma}^2_{M-1}\ (1+2\rho_{M-1}\hat{K}_M + \hat{K}^2_M) \qquad (64d)$$

which are used to determine $\rho_M = \alpha_M/\hat{\sigma}^2_M$ and then $\hat{K}_{M+1} = -\rho_M$ is used to initiate the Newton-Raphson search for the solution to

$$(N-M-1)\ \hat{K}^3_{M+1} + (N-2M-2)\ \rho_M\ \hat{K}^2_{M+1} - (N+M+1)\ \hat{K}_{M+1} - N\rho_M = 0 \qquad (65)$$

This algorithm has been found to be amenable to fixed point arithmetic and has been found to produce numerically stable estimates of the power spectral density. Furthermore, at the end of the recursive process, the data is immediately available for computing the spectral envelope which is extremely useful in certain low rate speech coding applications [3]. Typical results that are obtained using this procedure are shown in Fig. 2 for a voiced speech segment for a low pitched male speaker. The ML power measurements obtained from the peak-picking heuristic are indicated by the crosses at the top of the vertical bars. The spectral envelopes obtained by the exact and the approximate ML procedures are shown for comparison. Figure 3 illustrates typical results for a high-pitched female. While this represents a worst case condition with respect to the validity of the approximate solution, the spectral fit appears to be quite good. An example of the spectral envelopes obtained for an unvoiced speech segment is illustrated in Fig. 4.

A considerable amount of speech data has been examined graphically and it has been observed that for voiced speech the exact algorithm generally leads to wider formant bandwidths and reduced formant amplitudes. From (55) and (57) it is noted that the difference between the exact and approximate ML criteria is the term $-M \log (1-K^2_M)$ which has the effect of forcing the reflection coefficient towards zero, and causes the formant broadening. For unvoiced speech the approximate and the exact ML analyses result in

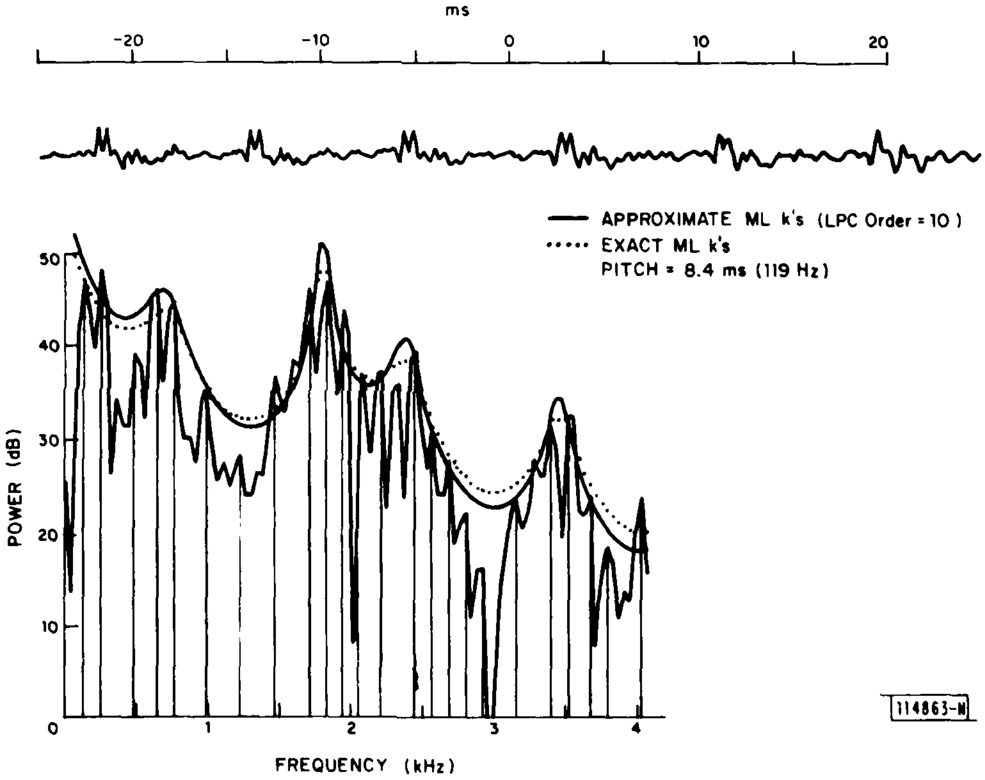

Fig. 2. Voiced speech spectral envelopes (male speaker pitch = 119 Hz).

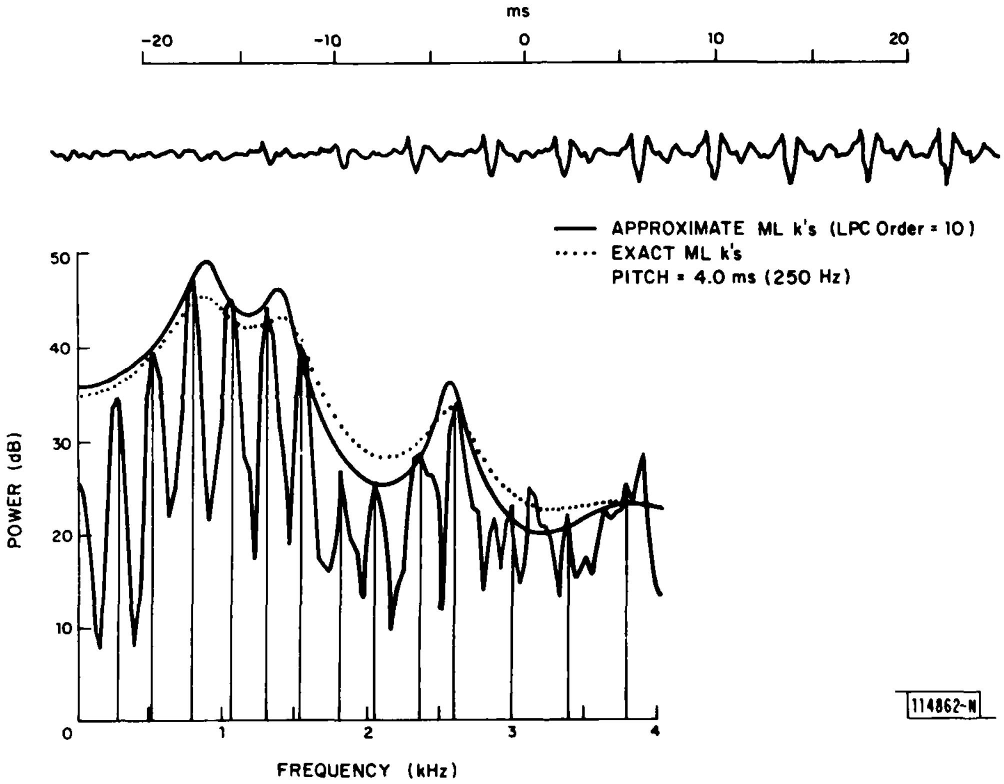

Fig. 3. Voiced speech spectral envelopes (female speaker pitch = 250 Hz).

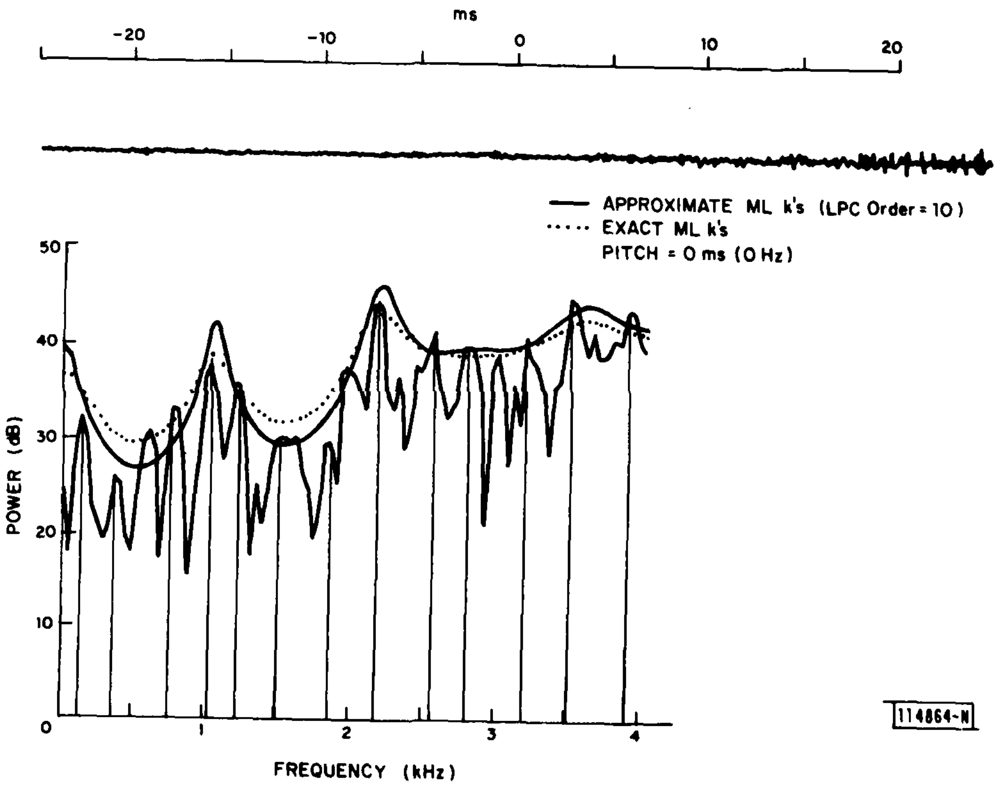

Fig. 4. Unvoiced speech spectral envelopes (female speaker).

spectral fits that appear to be essentially the same. It is interesting to note that even in the case of the low-pitched speaker, the curves in Fig. 2 show that the N>>M approximation is not valid. This suggests that for voiced speech it may be incorrect to interpret LPC spectral fits in terms of the Itakura-Saito spectral matching criterion. Rather it would appear that it would be more accurate to explain the LPC spectral fitting process in terms of spectral flattening as was done by Makhoul [8] and Markel and Gray [5].

In order to determine whether or not these numerical differences were perceptually detectable, a real-time speech analysis/synthesis vocoding system was developed using a tenth order spectral model over a 4 kHz audio bandwidth. The estimated spectral parameters were used in conjunction with a standard acoustic tube synthesizer and a high quality spectrally flattened filterbank synthesizer [9]. It was judged that the exact ML procedure resulted in synthetic speech that had the quality of having been too heavily smoothed. For the acoustic tube synthesizer, minor spectral distortions were produced occasionally when the approximate analysis algorithm was used. Although these distortions were not produced by the exact algorithm, the loss in speech naturalness caused by the excessive smoothing resulted in a preference for the approximate ML analysis system. These effects were less pronounced for female speakers. In all cases the differences were less noticeable when the filterbank synthesizer was used.

In another real-time evaluation the approximate pitch-directed frequency domain ML vocoder was compared with a standard autocorrelation LPC system. The quality and intelligibility of these systems were judged to be essentially equivalent. Therefore, from a perceptual point of view, there appears to be no advantage in using the frequency based spectral estimator that makes explicit use of the voiced and unvoiced speech models. It turns out, however, that the frequency domain implementation is particularly well suited to the development of new vocoding algorithms.

## V. SUMMARY AND CONCLUSIONS

The overall objective of this report was to try to determine whether or not better spectral estimators could be developed for speech by taking into account the periodic structure of the voiced speech sounds. As a starting point it was assumed that both voiced and unvoiced speech could be modelled as random processes. In the latter case the linear filter driven by white noise model not only leads to a convenient mathematical model, but also corresponds to the speech production mechanism. Voiced speech was modelled as a periodic random process, a random process having a periodic ensemble covariance function. Although this model has no relevance to the physiological mechanism by which voiced sounds are produced, mathematically it can be used to generate a class of spectra that have roughly the same properties as voiced speech spectra, and hence it was adopted as being suitable for the type of analysis to be undertaken. From this common mathematical framework, the random process could be expanded in terms of complex exponential basis functions, such that the statistical information was embedded in a set of expansion coefficients that were uncorrelated random variables. Assuming that the Gaussian model could be applied to the speech processes, the probability density function could be computed. The parameters of the underlying spectral model were embedded in a set of eigenvalues on which the pdf depended explicitly. Statistical estimates of these parameters were obtained using the maximum likelihood method and it was found that the nature of the resulting spectral fit was a highly nonlinear function of the logarithmic error between the measured and the model spectra. In fact this maximum likelihood spectral matching criterion was the same as that derived by Itakura and Saito [1] for the special case of unvoiced (aperiodic) sounds for which the linear filter was all-pole. The first important result of the analysis was to show that this spectral matching criterion was a general property of ML estimation and applied to aperiodic and to periodic processes and did not depend on a specific parametric spectral model. In particular, it did not need to be all-pole. Furthermore, although the Gaussian assumption was needed to derive the ML spectral matching criterion, it was no longer essential to the analysis, since any subsequent evaluation of the ML spectral estimates could be judged on the basis of the "goodness" of the spectral match.

The analysis was then specialized to the all-pole spectral envelope and solved exactly drawing heavily upon some recent work by Kay [2]. Hence-to-fore, solutions to the ML estimation problem were approximate requiring that the condition N>>M be satisfied where M was the all-pole model order and where, for unvoiced speech, N was the frame length. For voiced speech N was the pitch period or equivalently the number of harmonics in the audio bandwidth. Since a pitch of 300 Hz is not unusual, then for a 3600 Hz bandwidth there would be only 12 harmonics, hence, for a 10th order model it appeared that the results of the approximate ML analysis would not be valid for high-pitched speakers and perhaps was the reason for the poor performance of all-pole vocoders in this particular case.

In order to examine this conjecture in detail an algorithm was developed for determining the ML power measurements for speech using the pitch-directed peak picking heuristic developed by D. Paul [7]. Although only minor differences were observed when the approximate and the exact

ML procedures were applied to unvoiced speech, it was consistently noted that for voiced speech the exact analysis led to spectral estimates that widened the formant bandwidths while reducing the formant amplitudes. Evaluation using a real-time vocoder implementation revealed that the effect of the exact analysis was to produce synthetic speech that had the quality of having been smoothed excessively. The synthetic speech produced by the approximate analysis was judged to be more natural and was the preferred system, which shows that a strict implementation of the Itakura-Saito criterion is not perceptually desirable. Rather, the spectral flattening criterion that is implicit in LPC, seems to be preferable. The approximate ML system was then compared with a standard autocorrelation based LPC vocoder and the synthetic speech for both systems was judged to be essentially equivalent in quality and intelligibility.

Therefore, although a generalized and unifying theory for spectral analysis of aperiodic and periodic processes has successfully been developed, its application to narrowband speech coding has not led to significant perceptual improvements. That is not to say that the frequency domain formulation suggested by the ML procedure is not worthy of exploitation. On the contrary it has been crucial to the development of formant-based low rate systems [3] and the theory can be used as a basis for analyzing some of the time-domain pitch-adaptive algorithms that have already been developed [11]. Furthermore, the frequency domain implementation allows for the possibility of sampling the speech at a high rate, 10 kHz say, and analyzing the speech spectrum at any other lower rate simply by altering the cos(ω) table to reflect the desired folding frequency. This allows for the development of a new class of voicing adaptive split band vocoder algorithms, which, for a fixed model order (10 say) can adjust to a wider bandwidth (5 kHz say) during unvoiced speech and a lower bandwidth (3.4 kHz say) during voiced speech, and thereby result in "crisper" more intelligible speech at the same data rate. And finally it should be noted that care was taken to derive the likelihood function in such a way that all pitch-dependent terms were preserved so that it could be used as a pitch and voicing statistic.

## ACKNOWLEDGEMENTS

The author wishes to acknowledge the assistance received from Dr. D. Paul regarding the details of the SEEVOC analysis. Many stimulating conversations with Mr. J. Tierney and Dr. E. Hofstetter led to a better understanding of the properties of the frequency domain spectral fitting algorithm. Interaction with Dr. E. Hofstetter improved the details of the mathematical analysis leading to Equation (16).

# REFERENCES

[1] F. Itakura, and S. Saito, "A Statistical Method for Estimation of Speech Spectral Density and Formant Frequencies," Electronics and Communications in Japan 53-A, 36-43, (1970).

[2] S. Kay, "More Accurate Autoregressive Parameter and Spectral Estimates for Short Data Records," First ASSP Workshop on Spectral Estimation, McMaster University, Hamilton, Ontario, Canada, August 17-18, 1981, pp. 2.1.1-2.1.8.

[3] R. J. McAulay, "A Low-Rate Vocoder Based on an Adaptive Subband Formant Analysis," IEEE Intl. Conf. on Acoust., Speech and Signal Processing, Atlanta, Georgia, March 30 - April 1, 1981, pp. 28-31.

[4] H. Van Trees, Detection, Estimation and Modulation Theory, Part I (Wiley, New York, 1981).

[5] J. D. Markel, and A. H. Gray Jr., Linear Prediction of Speech (Springer-Verlag, New York, 1976).

[6] J. P. Burg, "Maximum Entropy Spectral Analysis," PhD Dissertation, Stanford University, 1975.

[7] D. B. Paul, "The Spectral Envelope Estimation Vocoder," IEEE Trans. Acoust., Speech and Signal Processing ASSP-29, 786-794 (1981).

[8] J. Makhoul, "Linear Prediction: A Tutorial Review," Proc. IEEE 561-580 (1975).

[9] B. Gold and J. Tierney, "Pitch-Induced Spectral Distortion in Channel Vocoders," Journal Acoust. Soc. Amer., 35, 730-731 (1963).

[10] R. J. McAulay, "Maximum Likelihood Spectral Estimation of Periodic Processes and Its Application to Narrowband Speech Coding," First ASSP Workshop on Spectral Estimation, McMaster University, Hamilton, Ontario, Canada, August 17-18, 1981, pp. 7.4.1-7.4.4.

[11] T. E. Tremain, J. W. Fussell, R. A. Dean, B. M. Abzug, M. D. Cowing, and P. W. Boundra Jr., "Implementation of Two Real-Time Narrowband Speech Algorithms," EASCON '78, Arlington, VA, September 25-27, 1978, pp. 698-708.

UNCLASSIFIED

SECURITY CLASSIFICATION OF THIS PAGE (When Data Entered)

## REPORT DOCUMENTATION PAGE

READ INSTRUCTIONS BEFORE COMPLETING FORM

1. REPORT NUMBER
ESD-TR-82-006

2. GOVT ACCESSION NO.
AD-A114 076

3. RECIPIENT'S CATALOG NUMBER

4. TITLE (and Subtitle)
Maximum Likelihood Spectral Estimation and Its Application to Narrowband Speech Coding

5. TYPE OF REPORT & PERIOD COVERED
Technical Report

6. PERFORMING ORG. REPORT NUMBER
Technical Report 602

7. AUTHOR(s)
Robert J. McAulay

8. CONTRACT OR GRANT NUMBER(s)
F19628-80-C-0002

9. PERFORMING ORGANIZATION NAME AND ADDRESS
Lincoln Laboratory, M.I.T.
P.O. Box 73
Lexington, MA 02173-0073

10. PROGRAM ELEMENT, PROJECT, TASK AREA & WORK UNIT NUMBERS
Program Element Nos.27417F, 28010F, 33401F
Project Nos.2283, 411L, 2264, 7820

11. CONTROLLING OFFICE NAME AND ADDRESS
Air Force Systems Command, USAF
Andrews AFB
Washington, DC 20331

12. REPORT DATE
5 March 1982

13. NUMBER OF PAGES
38

14. MONITORING AGENCY NAME & ADDRESS (if different from Controlling Office)
Electronic Systems Division
Hanscom AFB, MA 01731

15. SECURITY CLASS. (of this report)
Unclassified

15a. DECLASSIFICATION DOWNGRADING SCHEDULE

16. DISTRIBUTION STATEMENT (of this Report)

Approved for public release; distribution unlimited.

17. DISTRIBUTION STATEMENT (of the abstract entered in Block 20, if different from Report)

18. SUPPLEMENTARY NOTES

None

19. KEY WORDS (Continue on reverse side if necessary and identify by block number)

maximum likelihood
spectral estimation
periodic processes
aperiodic processes
all-pole modelling
narrowband speech coding
LPC

20. ABSTRACT (Continue on reverse side if necessary and identify by block number)

Using the maximum likelihood (ML) method the Itakura-Saito [1] spectral matching criterion is generalized to aperiodic and periodic processes having arbitrary model spectra. For the all-pole model, Kay's [2] covariance domain solution to the exact ML problem is cast into the spectral domain and used to obtain the exact solution for periodic processes. It is shown that if the number of independent power measurements greatly exceeds the model order, then the ML algorithm reduces to a pitch-directed, frequency domain version of Linear Predictive (LP) spectral analysis. Using a real-time vocoder based on the exact ML analysis revealed that, in contrast to standard LPC, the synthetic speech has the quality of being heavily smoothed. This suggests that it is generally incorrect to interpret LPC spectral matching in terms of the Itakura-Saito criterion.